Sociolinguística, corpora e educação

Sociolinguistics, corpora and education

Hadinei Ribeiro Batista*
Maria Cecília Mollica**

RESUMO

Os objetivos deste artigo são: (a) resenhar alguns trabalhos sobre construção de *corpora* para análise linguística, avaliando suas limitações para os estudos sociolinguísticos variacionistas; (b) discutir a necessidade do detalhamento de categorias sociodemográficas presentes em estudos variacionistas; e (c) propor um meio de construção de corpora, a partir de salas virtuais de aprendizagem, que viabilize o mapeamento e a descrição de variáveis linguísticas de forma mais transparente e que também possa contribuir para intervenções pedagógicas mais eficazes na educação básica.

Palavras-chave: *Sociolinguística, corpora, etnografia, ambientes virtuais, educação.*

* POSLING/UFRJ.
** UFRJ/CNPq/PPGCI/ FAPERJ.

ABSTRACT

The main objectives of this paper are: (a) review of some work on building corpora for linguistic analysis, evaluating their limitations for sociolinguistics studies, (b) discuss the necessity of detailing sociodemographic categories present in variation and change studies, and (c) propose ways of building corpora from virtual learning rooms, facilitating the work of mapping and describing linguistic variables in a transparent manner that can also contribute to more effective educational interventions in basic education.

Keywords: *sociolinguistics, corpora, ethnography, virtual environments, education.*

1. Introdução

Durante muitos anos, a Sociolinguística que produzimos no Brasil recebeu críticas procedentes por trabalhar com amostras de grandes proporções sem conhecer seus indivíduos. Os *corpora* existentes pré-estratificam em idade, escolaridade, sexo/gênero e em algumas outras variáveis dependentes, sem exatamente proceder a um trabalho etnográfico mais detalhado aos moldes de uma microanálise (HOLMES, 2008; DEAKIN; WAKEFIELD, 2014). Desta feita, há que se reconhecer que as análises de que dispomos na área têm pouco de sociolinguística e muito de linguística.

É de se notar ainda que a sociedade brasileira passou por profundas transformações. Além da erradicação da pobreza extrema, a nação conta hoje com a denominada Classe C, composta de mais de quase 30 milhões de brasileiros. Com acesso fácil ao crédito e a bens de consumo, o que anteriormente era privilégio da classe média e alta, a classe C passou a ter casa própria, possuir eletrodomésticos, como máquina de lavar, micro-ondas, televisão de última geração, sem falar de dispositivos mais modernos como computador e celulares de última geração. As viagens de avião tornaram-se realidade e a mesa farta já conta com iogurte, geleia, crepes, carne e uma sorte grande de itens oferecidos pela indústria alimentícia, cosmética e automobilística. A população tem acesso a entretenimento, pode cuidar-se

Revista Letras, Curitiba, n. 90, p. 221-231, jul./dez. 2014. Editora UFPR.
ISSN 0100-0888 (versão impressa); 2236-0999 (versão eletrônica)

em academias ao ar livre e sua expectativa de vida aumentou tanto quanto à das classes mais altas. De fato, é uma nova realidade brasileira, que quer serviços de qualidade, que exige dos governantes a garantia de preservação de conquistas e o que expressa o desejo de conquistar ainda mais.

Tendo em vista a ascensão da classe C, e agora das classes D e E, o Brasil é hoje a oitava economia mundial. No entanto, não consegue sair da categoria dos piores países em Educação no ranqueamento mundial. Por que esta discrepância? O que tem os linguistas a dizer aos economistas para explicar como os brasileiros não são letrados o suficientemente se já saíram da fronteira da miséria? Este é o desafio que se impõe à área; trata-se de tarefa que não pode esperar: incluir os brasileiros na cultura letrada, retirando-os das estatísticas de quase 10% de analfabetismo e 80% de analfabetismo funcional. Estaríamos nós, profissionais da Educação, equivocados?

Tudo leva a crer que a resposta, lamentavelmente, é afirmativa. Os bancos de dados de que dispomos não conhecem os sujeitos em seus detalhes, os quais os individualizam de forma que possamos conhecer de fato o perfil sociolinguístico dos membros de uma amostra para estudos sociolinguísticos. Sem um perfil identitário refinado, dificilmente vamos conhecer quais indicadores de bens de consumo influenciam para a inclusão na cultura letrada. Muitos estudos, tal como os reunidos por Roxo (2013), mostram a população conectada e o modo como a escola pode se valer das TICs (Tecnologias da Informação e Comunicação) para o trabalho dos multiletramentos.

Importa que as pesquisas e seus resultados nos apontem os indicadores que têm efeito positivo no letramento. Nesse sentido, as salas virtuais possibilitam conhecer exatamente a identidade dos indivíduos que estão por detrás das máquinas (MOLLICA; BATISTA; SILVA, a sair) para verificar a possível efetividade de traços na cultura letrada. Quanto maior o espectro de traços identitários, maior a chance de desvendar o mistério que interpõe um abismo entre as estatísticas econômicas e as de natureza educacional. Não sabemos, a priori, quais aspectos influenciam mais, razão por que a matriz que apresentamos vai aos mínimos detalhes. Pode ser que isso hoje não seja relevante; mas que, dada a celeridade de nossos tempos, o seja num futuro próximo. O dilema que se coloca é: conquistas de bens culturais asseguram um nível mais alto na lectoescrita?

Na pesquisa em andamento, buscamos discutir algumas dificuldades enfrentadas pela sociolinguística variacionista em relação ao controle de variáveis sociodemográficas. A amplitude dessas categorias em relação à diversidade de identidades sociais pelas quais a estrutura social se configura não permite um mapeamento transparente dos processos de variação e mudança linguística. Nossa proposta é detalhar essas variáveis com o

propósito de se aproximar mais das identidades dos sujeitos, almejando um encaixamento social mais satisfatório das variantes linguísticas.

Os métodos de construção de *corpus* ou banco de dados, seja através da Linguística de Corpus seja pelos mecanismos tradicionais da sociolinguística variacionista (entrevistas e narrativas orais), não se preocupam realmente com dados etnográficos. O controle de variáveis sociais não segue um padrão entre os diferentes *corpora*; como veremos, as variáveis sociodemográficas registradas, além de amplas, convergem categorias que não se correspondem mutuamente como *sexo/gênero*.

Um outro desafio que também buscamos superar em tal pesquisa é a construção de um banco de dados a partir de ambientes virtuais de aprendizagem, considerando a matriz etnográfica que ora colocamos em tese. As redes virtuais de interação, fonte rica de dados para pesquisas sobre a linguagem, de modo geral não se interessam pelo mapeamento de identidades sociais através dos procedimentos cadastrais. Muitas vezes, o cadastro de usuário possibilita a inserção de internautas fantasmas ou requer informações básicas, amplas e sem rigor sobre sua autenticidade. Tal procedimento dificulta a interface linguagem-sociedade e o cruzamento de variáveis sociais e comportamento linguístico que são cruciais para explicar fenômenos de variação e mudança linguística.

Sucintamente, redobramos nossa atenção visando à construção de uma matriz etnográfica para cadastro de usuários/informantes, capaz de resgatar detalhes de sua identidade social; também traçamos metas para a construção de um ambiente virtual de interação que sirva como fonte para a constituição de um banco de dados que se preste aos estudos sociolinguísticos variacionistas e a outras esferas de estudo no âmbito acadêmico. As seções a seguir expõem com detalhes os objetivos de nossa pesquisa, seus desafios, desdobramentos e seu estágio de evolução.

2. Desafios e limitações na construção de corpora para os estudos linguísticos

A Linguística de Corpus (LC) dedica-se à criação e à análise de *corpora*, objetivando colocar à disposição do analista da linguagem uma grande quantidade de dados. Acrescenta-se que essa área de estudo é uma abordagem cujo intento é possibilitar a investigação da linguagem e de todas as suas propriedades através de uma ampla coleção de amostras da língua, sejam estas escritas ou orais (SARDINHA, 2004; DASH, 2010).

As interfaces e diálogos da LC com outras áreas são também bastante amplas (OLIVEIRA, 2009) e incluem perspectivas no campo da sociolinguística. Baker (2010) argumenta que a LC pode oferecer à Socio-

Revista Letras, Curitiba, n. 90, p. 221-231, jul./dez. 2014. Editora UFPR.
ISSN 0100-0888 (versão impressa); 2236-0999 (versão eletrônica)

linguística não só um vasto volume de amostras da língua, mas também padrões de construção de *corpora* representativos de uma população. De fato, a LC possui um rigor metodológico envolvido na compilação de um *corpus*, cujos parâmetros objetivam garantir que a representatividade seja balanceada, adequada e que contemple, de forma segura e mais completa possível, a variedade linguística da população em análise (RASO, 2012; McENERY; WILSON, s.d.). A sociolinguística variacionista interessa-se por um grande volume de dados da linguagem espontânea. Mas, como relaciona língua e sociedade, depende de um efetivo controle de variáveis sociais. De modo geral, os *corpora*[1], compilados ou não conforme parâmetros da LC, não seguem um padrão no registro de fatores sociais e nem possuem uma matriz que sirva de referência. Dentre os parâmetros de compilação de *corpora* da LC, não há menção sobre o controle social dos informantes. Geralmente, registram-se fatores sociodemográficos amplos, como já é tradicional nos estudos sociolinguísticos.

Analisemos os fatores sociais registrados em três *corpora* do português: ALIP/IBORUNA, NURC-RJ e C-ORAL BRASIL.

1. **C-oral Brasil:** sexo, idade (faixa), escolaridade, ocupação e papel exercido na interação. (C-oral Brasil: p.56)

2. **NURC-RJ:** sexo, idade (em anos), naturalidade, naturalidade dos pais, escolaridade, área residencial. (<http://www.letras.ufrj.br/nurc-rj/>.)

3. **ALIP/IBORUNA:** sexo/gênero, idade (data nascimento), endereço completo, profissão, escolaridade, renda familiar.

(<http://www.iboruna.ibilce.unesp.br/arquivos_upload/arquivosAI/37/AI-011-CAS-2-FS.pdf>.)

1 Alguns corpora do português:
<http://www.nilc.icmc.usp.br/lacioweb/corpora.htm>.
<http://www2.lael.pucsp.br/corpora/bp/index.htm>.
<http://corpusbrasileiro.pucsp.br/cb/Inicial.html>.
<http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html>.
<http://www.letras.ufrj.br/nurc-rj/>.
<http://www.letras.ufrj.br/phpb-rj/>.
<http://www.linguateca.pt/>.
<http://www.c-oral-brasil.org>.
<http://www.iboruna.ibilce.unesp.br/interna.php?Link=corpo.php&corpo=36>.

Em relação ao fator *sexo,* o único corpus que registra o termo *gênero* é o ALIP, porém as opções são de *masculino* ou *feminino* ou *ambos.* Ora, *sexo* e *gênero* podem agrupar sujeitos distintos. O *gênero* envolve o sentimento de pertencer a uma categoria que não se limita às classificações objetivas da biologia: homem x mulher. A noção de *gênero* é menos estável e fixa do que a de *sexo* e isso influencia no comportamento linguístico do sujeito, pois a língua é também uma forma de impressão digital.

O registro da idade, também comum aos três *corpora*, é variável. O C-oral Brasil registra faixas etárias, incluindo o sujeito em intervalos de idades. O NURC-RJ registra a idade em anos e o ALIP/IBORUNA, a data de nascimento. Tal variação pode incorrer em impasses dependendo da pesquisa. Batista (2013), que trabalhou com dados do C-oral Brasil, não conseguiu confirmar a hipótese inicial sobre a mudança em progresso da interjeição *uai.* O autor verificou que as formas *uê/ué* eram usadas pelo grupo mais jovem e a forma *uai* pelos mais velhos. Porém, no intervalo de 26-40 anos, houve uma inversão desses usos e a falta de acesso à idade real dos informantes impossibilitou verificar se houve uma concentração de usos nas faixas mais próximas de 26 e nas mais próximas de 40. Vê-se, pois, que o registro da faixa etária não segue um padrão entre os diferentes *corpora*; assim, dependendo de sua notação, pode não satisfazer determinadas pesquisas.

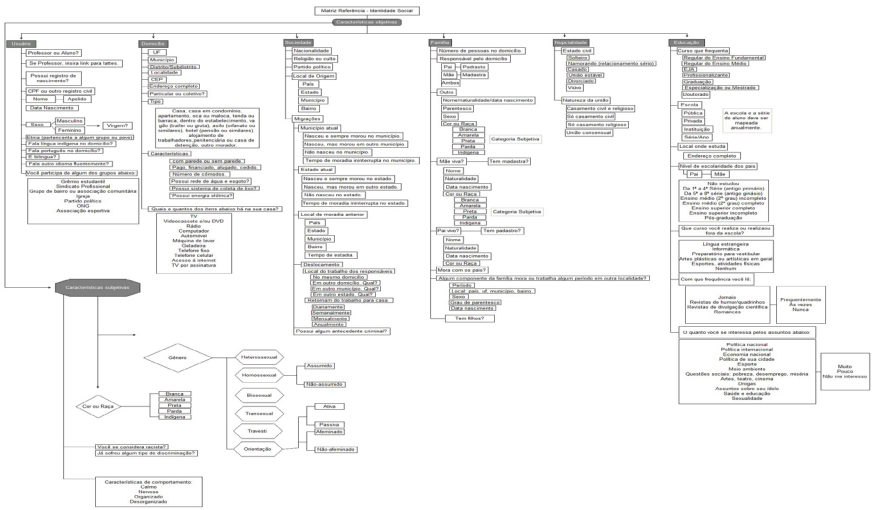
Outra diferença que também chama a atenção é a quantidade de variáveis sociais consideradas em cada *corpora*. Note que só o ALIP/IBORUNA registrou o fator *renda familiar.* Além disso, 5 ou 6 fatores sociais são suficientes para detalhar a identidade social de um informante? As categorias amplas empregadas tradicionalmente nos estudos sociolinguísticos quantitativos, embora tenham confirmado a variação no uso da linguagem, não são satisfatórias para explicar a porcentagem de informantes que se desviaram da regra geral. Somente categorias absolutas forneceriam um resultado convincente. A análise permite concluir uma tendência caso a porcentagem ou o peso relativo a favoreça, porém somente a apuração refinada de traços sociais dos indivíduos da amostra nos possibilitaria explicar os reais fatores que interferiram na porcentagem ou parcela de informantes que não realizaram o emprego. Se, em uma pesquisa, 70% dos homens utilizam uma forma e 30% não, é crucial justificar o fator social que impede a realização nesse grupo. A minoria dos informantes é sempre descartada porque as categorias que se leva em conta não são detalhadas.

Revista Letras, Curitiba, n. 90, p. 221-231, jul./dez. 2014. Editora UFPR.
ISSN 0100-0888 (versão impressa); 2236-0999 (versão eletrônica)

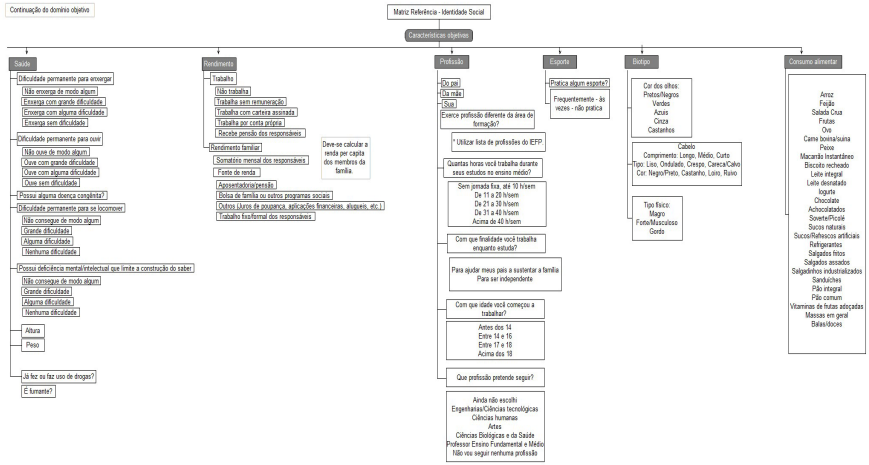
3. Sociolinguística e etnografia: detalhamento de categorias sociodemográficas

Batista e Mollica (2014)[2] dedicaram-se à construção de matrizes identitárias a partir de dados/traços sociais de diferentes domínios. A proposta seguiu parâmetros do Censo Demográfico do IBGE. O objetivo foi elencar um conjunto vasto de traços identitários, agrupados em diferentes categorias (sociedade, família, saúde, nupcialidade etc.). As matrizes, embora tenham sido desenvolvidas com o intuito de possibilitar um mapeamento de usuários dos ambientes virtuais previstos pelo projeto, servirá também de base para pesquisadores que tenham interesse em tais registros em suas pesquisas, principalmente aqueles envolvidos em construção de *corpora.* Vários autores que tratam do mapeamento da identidade social de um indivíduo (HALL, 1999...) pontuam, como critério para particularizar um indivíduo em seu grupo social, um conjunto de traços. Desse conjunto, alguns são semelhantes a outros indivíduos, mas, na medida em que se amplia esse repertório, a tendência é se aproximar, cada vez mais, da identidade real e individual de uma pessoa. Nesse sentido, optou-se por elencar o maior número de fatores sociais possíveis para fins de investigar, em maior detalhe, a influência de traços específicos no comportamento linguístico de um falante, o que justifica a utilização de uma matriz identitária como proposta pelos autores supramencionados.

Abaixo, copiamos uma das matrizes propostas pelos autores.



2 O artigo sobre a construção da matriz está em processo de publicação e pode ser baixando em: <http://goo.gl/95EfQn>. Os autores também já desenvolveram matrizes identitárias para diferentes níveis da educação básica. O texto está disponível em: <http://goo.gl/T4a98e>.



4. Ambientes virtuais, educação e construção de corpora

Outro subtema da pesquisa em andamento é a construção de salas virtuais de aprendizagem, já mencionadas em outro trabalho (BATISTA; MOLLICA, 2014)[3]. Observa-se, atualmente, uma lacuna em relação à educação a distância (EAD). As EADs normalmente envolvem ambientes virtuais que refletem a forma de organização da educação tradicional. A única diferença é que a interação entre docentes e discentes se dá a distância. Apesar da globalização possibilitada pelas tecnologias de informação, alunos e professores da educação básica ainda mantêm laços interacionais limitados ao contexto físico da escola. Nossa proposta de salas virtuais de aprendizagem tem o caráter inovador de oferecer a educadores e educandos um recurso complementar à educação tradicional, visando um acesso a um conhecimento mais global, democrático e flexível.

Não é novidade a quantidade variada de plataformas virtuais de aprendizagem. Os ambientes que funcionam conforme as políticas da EAD oferecem cursos a distância, com obrigatoriedade de encontros presenciais determinados em leis que regulam essa modalidade de ensino. Há outros ambientes que proveem um meio de ensino mais flexível. Vários cursinhos pré-vestibulares contam com um ambiente virtual em que professores ficam à disposição dos alunos durante certo período do dia. Isso favorece aqueles educandos que preferem estudar no conforto do lar e no horário que estão mais dispostos. Mesmo assim, essas ferramentas são específicas para alunos

3 Disponível em: <http://goo.gl/T4a98e>.

Revista Letras, Curitiba, n. 90, p. 221-231, jul./dez. 2014. Editora UFPR.
ISSN 0100-0888 (versão impressa); 2236-0999 (versão eletrônica)

matriculados nos cursos e, em geral, o acesso à tal plataforma requer apenas usuário, senha e número de matrícula. Os aprendizes continuam limitados a interações com professores do cursinho e não têm sua identidade social rigorosamente mapeada para possíveis intervenções pedagógicas.

As salas virtuais que ora propomos assemelham-se à proposta dos cursinhos acima referida, porém de forma gratuita e com alcance bem mais amplo. A partir da matriz etnográfica prevista em nossa pesquisa, alunos e professores do sistema básico de ensino seriam cadastrados, obedecendo-se a um rigoroso controle em relação à precisão e à autenticidade das informações requeridas. Assim, todos os usuários, para ter acesso ao portal, teriam sua identidade social filtrada pelas telas do cadastro. Para evitar fadiga na entrada de dados, pois a matriz é muito ampla, alguns traços, já considerados fundamentais na investigação sobre variação linguística, serão de entrada obrigatória; o restante será introduzido aos poucos em telas dinâmicas, que favoreçam a inserção dos traços identitários. Embora nenhum usuário venha a ser identificado na plataforma, uma vez que sua privacidade é legalmente garantida, os fatores sociais cadastrados serão disponibilizados para pesquisas em centros acadêmicos.

As salas virtuais de aprendizagem seriam categorizadas conforme as disciplinas previstas nos PCNs. O aluno acessaria, a seu tempo e lugar, a sala de seu interesse. Nessa sala, ele seria informado sobre alunos e professores disponíveis para interação instantânea. A interação teria modo escrito ou *face-time.* Através desse canal, aprendizes de qualquer lugar do Brasil poderiam tirar dúvidas, fazer trabalhos e receber orientações diversas que promovam seu acesso ao conhecimento de maneira autônoma, produtiva e eficaz. Trata-se, na verdade, de uma ferramenta complementar capaz de ampliar os limites físicos da sala de aula, projetando o aluno para uma sala virtual com uma diversidade de escolhas e de formas de aprender condizentes com as potencialidades proporcionadas pela mídia digital. A ampliação do acesso a diversos outros professores e alunos expandiria os laços sociais dos educandos, criando condições propícias para o atendimento de suas necessidades individuais, conforme previsto na Declaração de Salamanca (1994) sobre as práticas educativas especiais.

Além de contribuir para uma educação mais igualitária, global e democrática, os ambientes, por serem monitorados etnograficamente, colocariam à disposição de pesquisadores um conjunto de dados para estudos acadêmicos. Intervenções pedagógicas teriam condições de ser mais pontuais, já que seria possível filtrar grupos de alunos e de professores que necessitam de capacitação/orientação pedagógica especial. A filtragem forneceria as variáveis sociais dos sujeitos a ponto de revelar as implicações desses fatores no seu desempenho e na construção autônoma do seu conhecimento. Isso

abriria um leque de opções para a implementação de políticas públicas para elevar a qualidade do ensino do nosso país.

Voltando para os estudos sobre a linguagem, as interações, sejam escritas sejam orais, constituiriam uma amostragem bastante vasta do uso natural da língua, propiciando estudos em diferentes níveis gramaticais, incluindo os estudos que relacionam língua e sociedade pelo acesso às informações etnográficas cadastradas no ambiente. Salienta-se que dados virtuais em estudos linguísticos sofrem inúmeras restrições. Baker (2010) cita um estudo de King (2009) em que o autor construiu e analisou um *corpus* com dados extraídos de salas de bate-papo do inglês americano e australiano, cujos participantes eram homossexuais. O estudo pretendia investigar se havia um interesse dos participantes em alterar os nomes dos internautas com emprego de diminutivos ou com estratégias de "feminização" desses nomes. Apesar da sala de bate-papo ser reservada para o público gay, o acesso era livre e heterossexuais podiam ter feito parte de tal amostra. Além desse entrave, o acesso a vários ambientes virtuais não possui um controle rigoroso de informações pessoais do internauta e nem é solicitado a este autorização para uso de seus dados em pesquisas acadêmicas. De acordo com Baker, King precisou enviar um e-mail para cada internauta, requerendo autorização para fazer uso das informações.

Tais impasses mostram a emergência de uma metodologia capaz de controlar etnograficamente o fluxo de informações na internet. As salas virtuais que nossa pesquisa pretende construir vão ao encontro desse objetivo, superando as lacunas sobre o controle identitário dos usuários virtuais e a falta de globalização da educação pública nacional.

5. Conclusões

O presente texto discutiu a importância de um maior detalhamento do perfil social de sujeitos envolvidos em pesquisa acadêmica e na construção de *corpora.* A matriz de referência ora proposta aponta um mecanismo eficiente na coleta de traços identitários de usuários através de um plano de cadastro para ambientes virtuais que priorizam redes de interação instantânea.

De fato, a construção de *corpora* a partir de dados virtuais espontâneos é um desafio recente. O projeto em tese, pelo que foi exposto, almeja disponibilizar uma ferramenta capaz de subsidiar a educação básica e coletar um conjunto vasto de informações para pesquisas sobre a linguagem e sobre o processo de ensino-aprendizagem, já que restringe as interações ao contexto educacional.

Revista Letras, Curitiba, n. 90, p. 221-231, jul./dez. 2014. Editora UFPR.
ISSN 0100-0888 (versão impressa); 2236-0999 (versão eletrônica)

REFERÊNCIAS

BAKER, P. Sociolinguistics and Corpus Linguistics. Edinburgh: Edinburgh University Press. 2010.

BATISTA, Hadinei Ribeiro. Uai: estudo de uma interjeição do português brasileiro. 117fls. Dissertação (Mestrado) – FALE/UFMG, Belo Horizonte, 2013.

BERBER SARDINHA, T. Lingüística de Corpus. São Paulo: Manole. 2004

_____. Lingüística de Corpus: uma entrevista com Tony Berber Sardinha. Revista Virtual de Estudos da Linguagem – ReVEL, vol. 2, n. 3, ago. 2004.

DASH, Niladri Sekhar. Corpus Linguistics: A General Introduction. Workshop on Corpus Normalization, LDCIL, CIIL, 2010, Mysore, India. Proceedings... Disponível em: < http://www.ldcil.org/download/Corpus%20Linguistics.pdf>. Acesso em 30 out. 2014.

DEAKIN & WAKEFIELD. Skype interviewing: reflections of two PhD researchers. Qualitative research, 2014.

HOLMES, S. Methodological and Ethical considerations in designing an Internet study of quality of life: A discussion paper. International Journel of Nursing studies, 2009.

KENDALL, Tyler. Corpora from a sociolinguistic perspective (Corpora sob uma perspectiva sociolinguística). *Revista Brasileira de Linguística Aplicada*, 11.2, 2011. p. 361-389.

McENERY, T., WILSON, A. Corpus Linguistics: what is a corpus and what is in it?. Disponível: <http://www.lancaster.ac.uk/fss/courses/ling/corpus/Corpus2/2FRA1.HTM> Acesso em: 13 jan. 2014.

MOLLICA, Mara Cecilia; BATISTA, Hadinei; SILVA, Cynthia (orgs). Sujeitos em ambientes virtuais. Rio de janeiro: Editora Mauad, a sair.

OLIVEIRA, Lúcia Pacheco. Linguística de corpus: teoria, interfaces e aplicações. Matraga, Rio de Janeiro, v. 16, n. 24, jan/jun/2009.

RASO, Tommaso; MELLO, Heliana Ribeiro de. C-ORAL-BRASIL I: corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012. 332 p.

ROJO, Roxane (org.) Escol@ conectada: os multiletramentos e as TICs. São Paulo: Parábola Editorial, 2013.

Submetido em: 25/04/2014

Aceito em: 16/10/2014